Credit Card Fraud Detection

By Aniket Khodankar, Arshiya Begum, Sarath Nandiminti

OBJECTIVE

Finex is a leading financial service provider based out of Florida, US. It offers a wide range of products and business services to the customers through different channels like in-person banking, ATMs and online banking.

In recent times, the number of fraud transactions has increased drastically due to which the company has been facing a lot of challenges.

For banking companies like Finex, retaining high profitable customers is the most important business goal. With the rise in digital payment channels, banking fraud, however, poses a significant threat to this goal for many banks.

As a consulting company hired by Finex, our task was to identify the root cause of this issue of unauthorized transactions on credit card/debit card and recommend ways to mitigate this problem.

PROBLEM STATEMENT

- Fraudsters steal credit card information using skimmers in ATM/POS terminals and make unauthorized transactions.
- The Federal Trade Commission estimates that 10 million people are victims of credit card theft each year.
- Credit card companies lose close to $50 billion per year to fraud.
- These costs ‘trickle down’ to higher interest rates and fees for all consumers.
- All cardholders pay for credit card fraud losses.

  ● Victims spend time and money to repair the damage.

  ● Credit card issuers charge higher fees and interest rates to cover their losses.

Bank's Perspective:

- Average number of transactions per month = 77183
- Average number of fraudulent transaction per month = 402
- Average amount per fraud transactions = ~$530

Cost incurred due to fraudulent transactions earlier = 530 * 402 = 213060

Cost incurred per month after the model is built and deployed = 2723

Final savings = Cost incurred before - Cost incurred after

= 213060 - 2723 = 210337 (98.7% savings!!)

PROJECT GOALS

The primary goal of our project was to reduce the occurrence of credit card fraud at Finex by Implementing a robust fraud detection model.

We aimed to achieve two key outcomes: identifying the root causes of unauthorized transactions and recommending actionable solutions to minimize these transactions moving forward.

Ultimetly our goal was to design a model that can predict fraudulent transactions with a high degree of accuracy.

APPENDIX - PROBLEM SOLVING METHODOLOGY

- **Problem Solving Methodology**

The approach for this project has been designed to follow the **CRISP DM Framework.** The various stages of the framework are represented below in a sequential flow:

APPENDIX: DATA SOURCES

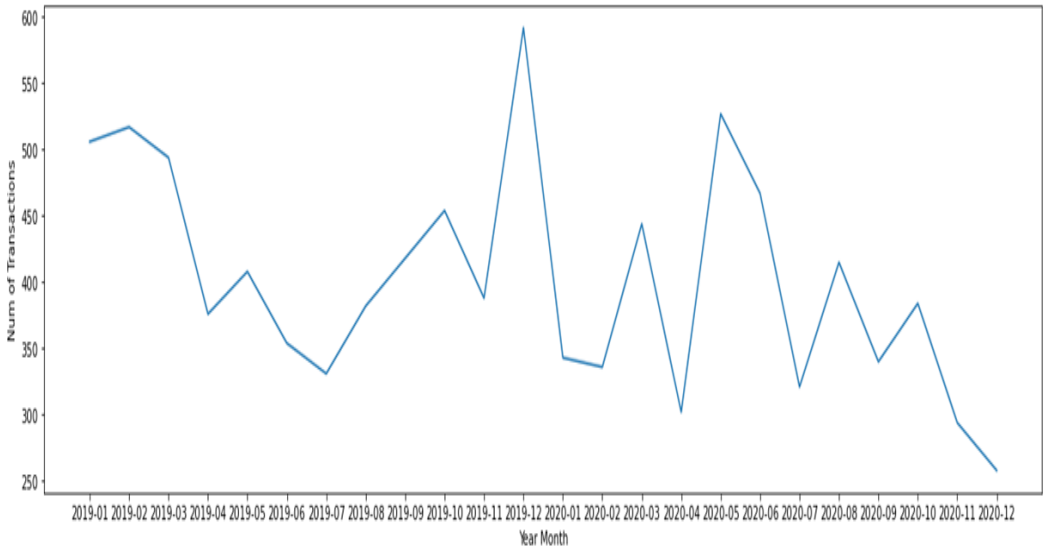
We have a simulated credit card transaction data set containing legitimate and fraudulent transactions from 1 January 2019 to 31 December 2020. Source: Kaggle

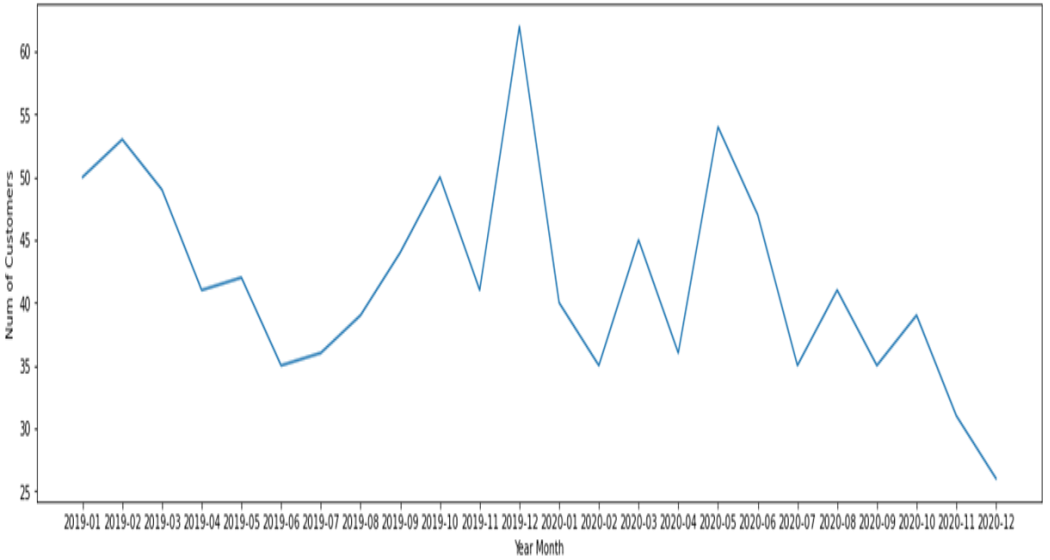
It covers credit cards of 1,000 customers doing transactions with a pool of ~700 merchants.

A transactional data usually contains the information of the customer, merchant, location variables, transaction date time, transaction amount and whether it is a fraud or not.

YEAR MONTH VS FRAUD TRANSACTIONS AND CUSTOMERS

- The number of fraudulent transactions vary from 250 to 600. Fraudulent transactions are distributed across all months which shows fraud is a continuous behaviour.
- Interestingly, in Dec -2020, there is a peak of overall transactions, whereas fraudulent transactions show a dip.
- Fraudulent transactions are repeated for the same customers' credit cards; for example, 500 fraudulent transactions were made with the same 50 customer's credit cards in January 2019. **Mostly, more than one fraudulent transaction occurred with the same customer's credit card.**





MODEL RESULTS – DECISION TREE

```
Test Results
[[1247825    1496]
 [     29    6449]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1249321
           1       0.81      1.00      0.89      6478

    accuracy                           1.00   1255799
   macro avg       0.91      1.00      0.95   1255799
weighted avg       1.00      1.00      1.00   1255799

 Train Results
 [[1216105    1522]
  [      0 1217627]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1217627
           1       1.00      1.00      1.00   1217627

    accuracy                           1.00   2435254
   macro avg       1.00      1.00      1.00   2435254
weighted avg       1.00      1.00      1.00   2435254
```

- The decision tree model that has been built has good precision and recall in the train, but the precision is hugely dropped in the test data. This indicates the overfitting problem. So, we cannot use this model for the prediction of fraud transactions.

Note: Precision and recall need to be only evaluated on the minor class that is Fraud (1) in our case.

MODEL RESULTS – RANDOM FOREST

- The random forest model has good precision and recall in the train and similar precision and recall in the test data. So, this model will be good fit in the prediction.

Train Results

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  50 out of  50 | elapsed:   36.3s finished
```

```
[[1217541      86]
 [      0 1217627]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1217627
           1       1.00      1.00      1.00   1217627

    accuracy                           1.00   2435254
   macro avg       1.00      1.00      1.00   2435254
weighted avg       1.00      1.00      1.00   2435254
```

Test Results

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  50 out of  50 | elapsed:   20.6s finished
```

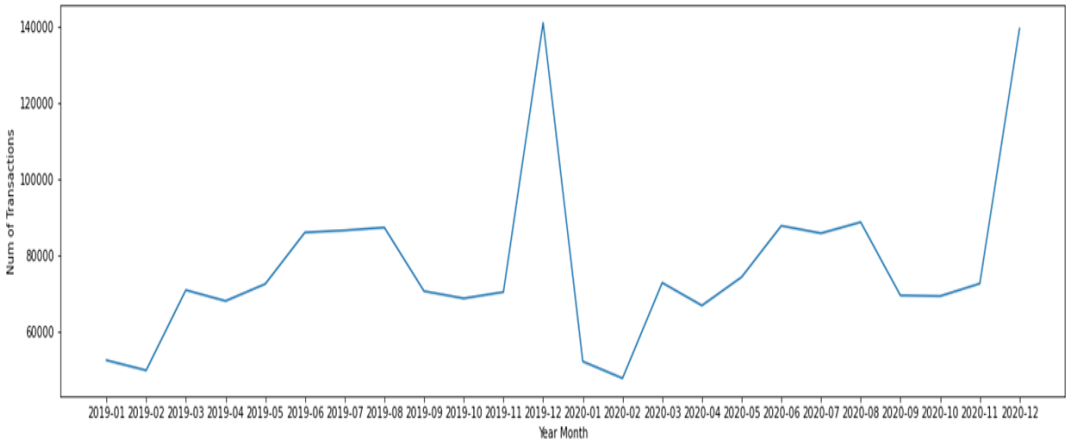
```
[[1249224      97]
 [     46    6432]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1249321
           1       0.99      0.99      0.99      6478

    accuracy                           1.00   1255799
   macro avg       0.99      1.00      0.99   1255799
weighted avg       1.00      1.00      1.00   1255799
```

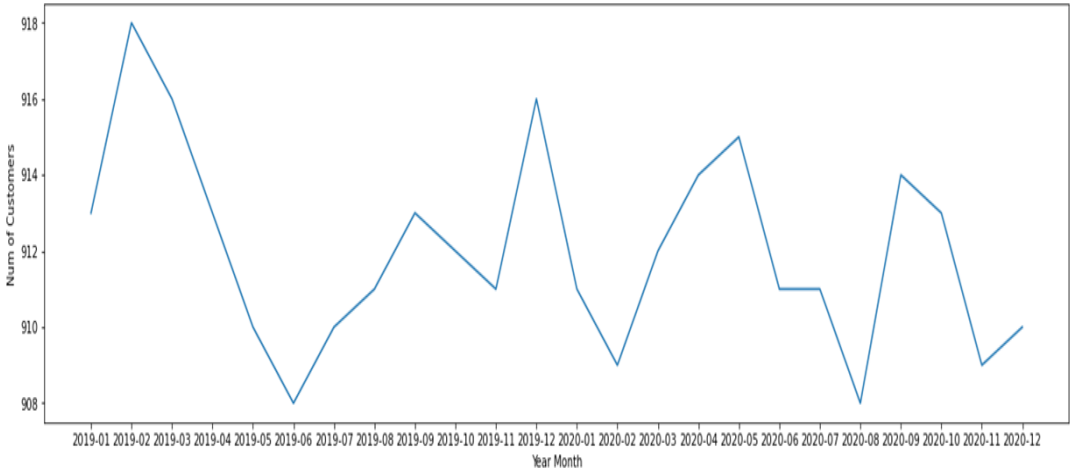
KEY FINDINGS
YEAR MONTH VS TRANSACTIONS AND CUSTOMERS

The timeline plot shows the number of transactions and the number of customers who made transactions in a given year month.

The number of transactions by ~900 customers ranges from 60k to 140k and changes every month.

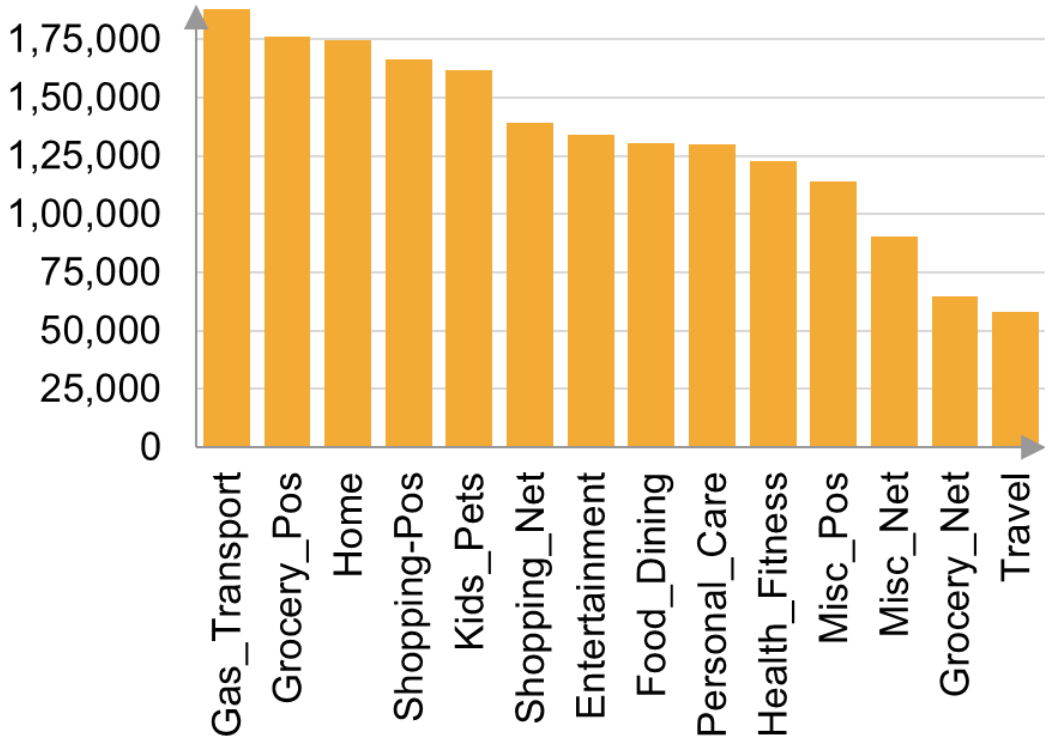
The number of customers did not vary a lot. It shows that the **same number of customers make multiple transactions, which is also a seasonal behavior**. We see a peak in the month of December.





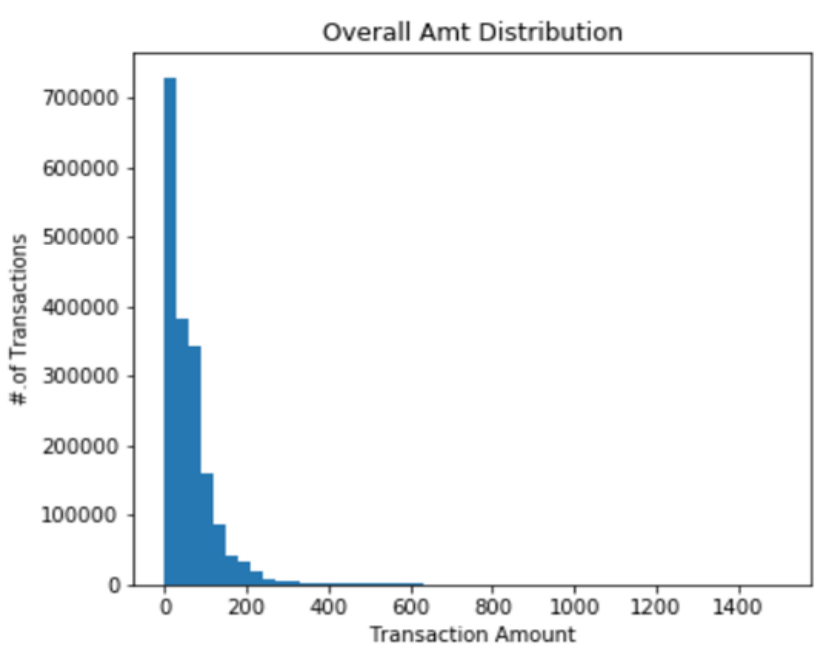
CATEGORICAL VARIABLE – MERCHANT CATEGORY

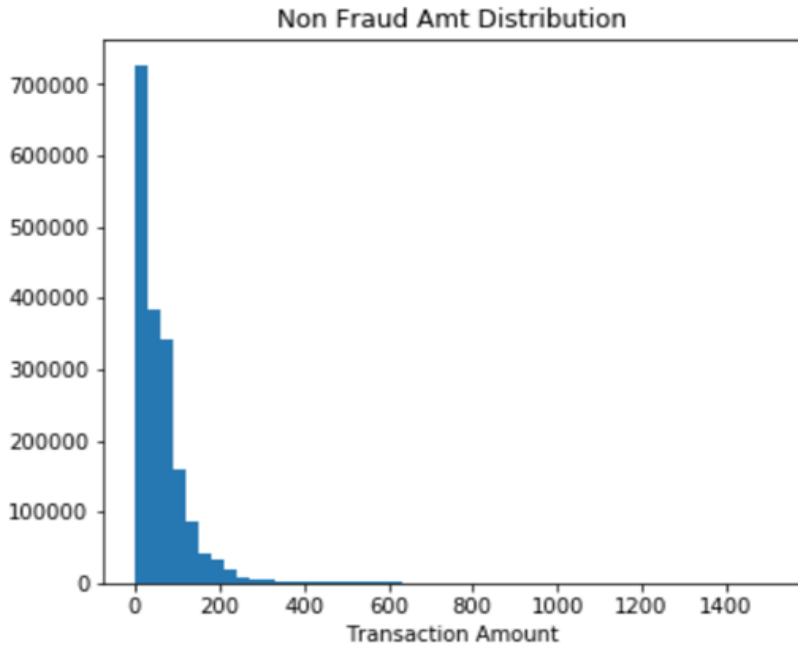
- The number of transactions varies between the categories (gas_transport, grocery_pos …. travel).
- **The fraud percentage (Ref percent_grp in table) also shows different levels based on the categories.**
- The category ‘Health_Fitness’ has a low percentage of fraudulent transactions, whereas the category ‘Shopping_Net’ has a high percentage of fraudulent transactions.

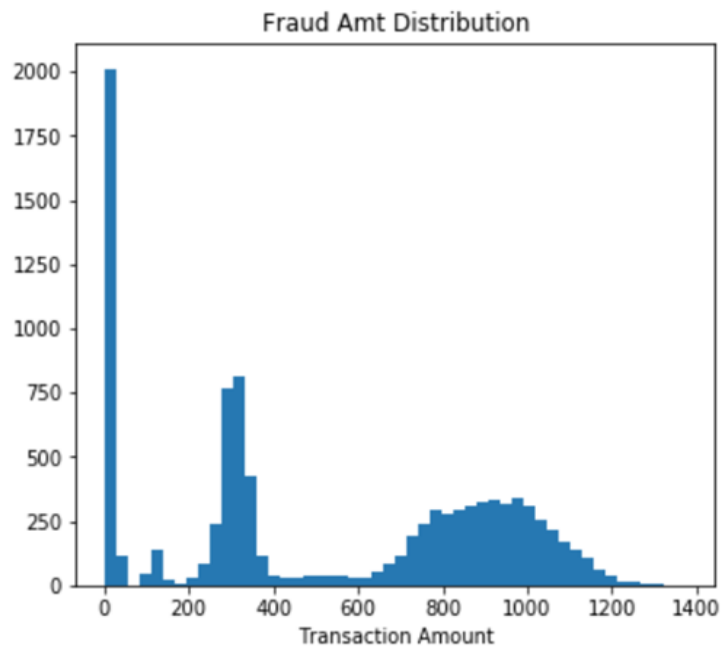


CATEGORY	IS_FRAUD	COUNT	CATEGORY_COUNT	PERCENT	PERCENT_GRP
Health_Fitness	1	185	122553	6.615925	0.150955
Home	1	265	174560	9.472067	0.151032
Food_Dining	1	205	130729	7.057300	0.156813
Kids_Pets	1	304	161727	8.730702	0.187971
Entertainment	1	292	134118	7.240252	0.217719
Personal_Care	1	290	130085	7.022534	0.222931
Travel	1	156	57956	3.128708	0.269170
Grocery_Net	1	175	64878	3.502387	0.269737
Misc_Pos	1	322	114229	6.166561	0.281890
Gas_Transport	1	772	188029	10.150594	0.410575
Shopping_Pos	1	1056	166463	8.986371	0.634375
Grocery_Pos	1	2228	176191	9.511529	1.264537
Misc_Net	1	1182	90654	4.893883	1.303859
Shopping_Net	1	2219	139322	7.521186	1.592713

AMOUNT DISTRIBUTION







- The distribution of 'Amt' for fraud transactions is quite different from the Overall Amt Distributions.
- **Fraudulent transactions are concentrated in ranges of [$1 - $10] , [$200 - $400] and [$600 - $1,200] bill values.** Fraudsters are mostly focusing on mid value range of Transaction Bill Value and not on very high-value Bill Value transactions as in 'Non Fraud' Transactions which is between $1,500 to $ 30,000].

*Note: Transactions Amt great then 1500 is removed in Overall and Non Fraud Plots for Just Visualizing the data distribution, which is 0.1% of the overall transactions.

CONCLUSION

In conclusion, our project successfully developed a fraud detection transactions with high accuracy but also saves Fines a substantial amount in fraud-related costs. This has significant implications for the banking industry as a whole, where such models can be deployed to reduce fraud and safeguard customer trust.

We believe that further refinement of these models and integration with real time detection systems with enhance fraud prevention efforts even more.

THANK YOU